

DOCUMENT RESUME

ED 129 576

SE 020 858

AUTHOR Biggs, Jon S.; Whitehead, Gary R.
TITLE An Empirical Study of the Pseudo-Paradoxical Effect on Correlation of Extreme Univariate Selection in the Three-Variable Case.
PUB DATE Apr 76
NOTE 14p.; Paper presented at the annual meeting of the American Educational Research Association (San Francisco, California, April 19-23, 1976); Not available in hard copy due to marginal legibility of original document.
EDRS PRICE MF-\$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Correlation; *Data Analysis; *Educational Testing; Elementary Secondary Education; *Research; Sampling; Standardized Tests; Statistical Analysis; *Statistical Bias; Tests
IDENTIFIERS Research Reports

ABSTRACT

Five three-variable combinations were obtained from a variety of standardized test scores of 331 male children. These five groups were treated as unrestricted populations of the case III variety. Extreme univariate selection was produced in each group allowing for tests of differences between the correlation values of the two incidental variables in the restricted and unrestricted groups and between the predicted correlation of incidental variables in the restricted sample and the actual value. Results indicated that the case III correction formula is appropriate under conditions of extreme univariate selection and that restriction of range does not invariably decrease correlations. (Author/SD)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

ED129576

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

13.15

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

AN EMPIRICAL STUDY OF THE PSEUDO-PARADOXICAL EFFECT
ON CORRELATION OF EXTREME UNIVARIATE SELECTION
IN THE THREE-VARIABLE CASE

Jon S. Biggs

State University of New York

College at Fredonia

Gary R. Whitehead

Mississippi State University

BEST COPY AVAILABLE

Paper Presented at 1976 AERA Convention, held in
San Francisco, California, April 21, 1976.

One of the traditional problems of measurement in education is accurately inferring parameters for values obtained from a restricted sample. Educational researchers frequently encounter situations where representative sampling in the collection of data is either inconvenient or altogether impossible. When such a situation occurs within the correlation model, investigators typically resort to biased or selected samples. Values obtained from these samples are then often presented with the warning of possible inaccuracy in the direction of underestimation, due to the restriction of range within the data. These warnings at first appear somewhat peculiar in view of the fact that formulas developed to correct sample statistics for the effects of restriction are available. However, in examining the literature relevant to these correction formulas one finds a dearth of empirical research and for this reason it is not difficult to understand why they are so infrequently employed.

Pearson (1903) addressed this problem of inferring correlational values, developing solutions for estimating parameters in three different situations. The situation of interest in this paper is the three-variable case, or as it is frequently labeled "case III". In the three-variable case the value of interest, the parameter to be inferred, is the correlation between two variables y and z in an unrestricted population (R_{yz}). The functional values are the correlation of y and z in the restricted sample (r_{yz}), the correlations in the restricted sample between a selection variable x and both incidental variables y and z (r_{xy} and r_{xz} , respectively), and the variance of x in the restricted and unrestricted population. The correction of r_{yz} for restriction of range

given by Pearson (1903) is found in (1).

$$R'_{yz} = \frac{r_{yz} + Kr_{xy}r_{xz}}{\sqrt{1 + Kr^2_{xy}} \sqrt{1 + Kr^2_{xz}}} \quad \text{where} \quad (1)$$

- R'_{yz} = the predicted correlational value of y and z in the unrestricted population,
 r_{yz} = the correlational value of y and z in the restricted sample,
 r_{xy} = the correlational value of x and y in the restricted sample,
 r_{xz} = the correlational value of x and z in the restricted sample,
 $\sigma^2_{x'}$ = the variance in the unrestricted population,
 σ^2_x = the variance in the restricted sample, and
 $K = \frac{\sigma^2_{x'}}{\sigma^2_x} - 1$

Formula (1) would estimate the correlation of variables y and z, when selection had been made on the basis of the third variable x. The derivation of the correction for restriction of range formula is based on the following assumptions (Lord & Novick, 1968): the regressions of y and z on x are linear, the variance of y given x, the variance of z given x, and the covariance of z and y given x, do not depend on x. Traditional statistical theory maintains that when these assumptions are met the use of the correction formula will yield a correlational value greater than that found in the restricted sample and more nearly like that which exists in the unrestricted population, and moreover, that the formula should always be employed under these conditions to avoid underestimates of the actual value of R_{yz} (Peters & Van Voorhis, 1940; Thorndike, 1949; Gulliksen, 1950; Gullford, 1956; Rydberg, 1964).

To date only three sources have questioned the accuracy of the correction formula and the validity of the assertion that restriction of range always results in a decrease in R_{yz} . Lord and Novick (1968) recommended a more cautious attitude toward the correction formula, especially in any application where the selection ratio is more than 1.40.

Levin (1972) employing formula (1) where K is defined as $\frac{\sigma^2_{x'} - 1}{\sigma^2_x}$ pointed out that by definition $\sigma^2_{x'} > \sigma^2_x$ and K is therefore positive. Also since (1) is symmetrical in r_{xz} and r_{xy} , the investigation of R'_{yz} as a function of either of these correlations is sufficient.

As r_{xz} in (1) approaches zero, R'_{yz} tends to (2).

$$R'_{yz} = \frac{r_{yz}}{\sqrt{1 + Kr^2_{xy}}} \quad (2)$$

Since the denominator is larger than unity, $R'_{yz} < r_{yz}$, i.e., the correlation in the unrestricted population is less than the correlation in the restricted sample. This is a reversal of the usually obtained inequality and contradictory to traditional theory. Therefore, according to this derivation there exists a range of values of r_{xz} (for fixed values of all other parameters in (1)) where explicit selection on x increases the correlation between the variables y and z . Using artificial data, Levin (1972) demonstrated that this inequality holds, not only for sufficiently small values of r_{xz} , but also for sufficiently large values of r_{xz} and he referred to this finding as the "pseudo-paradoxical" effects of range restriction.

Elshout and Roe (1973) in a paper concerned with the inferential accuracy of correction formulas stated that selection can make r_{yz} rise above R_{yz} , whenever the difference between R_{xy} and R_{xz} is great enough.

and R_{yz} is above a certain critical value. They define this critical value as the value of R_{yz} for which given R_{xy} , R_{xz} , and σ^2_x , selection makes no difference at all. The critical point is defined in (3).

$$R_{yz} = \frac{R_{xy}R_{xz}(1 - \sigma^2_x)}{1 - \sqrt{1 - R_{xy}^2(1 - \sigma^2_x)} \sqrt{1 - R_{xz}^2(1 - \sigma^2_x)}} \quad (3)$$

It is important to note that the work of Levin (1972) and that of Elshout and Roe (1973) was theoretical, i.e., only artificial data were employed.

Despite the obvious importance of this problem, very little work has been done to check the accuracy of the correction formula (Lord & Novick, 1968). A study was reported by Thorndike (1949). Another was done by Peters and Van Voorhis (1940), and another by Rydberg (1963). Each of these writers took a rather positive attitude toward the correction formula, insisting that corrected values are more valid as measures of relationship in the unrestricted population than the uncorrected values. However, as Lord and Novick (1968) noted, none of these studies involved extreme selection.

The problem involved in this study was twofold. First, an attempt was made to determine whether or not correlational values obtained from real data increase or remain unchanged as a result of restriction of range under those conditions outlined by Levin (1972) and Elshout and Roe (1973). Secondly, an attempt was made to assess the accuracy of the case III correction formula under the conditions of extreme univariate selection. Extreme univariate selection was defined as a selection ratio of 1.5 or greater.

Method and Results

Data were obtained from the records of a child development clinic. Data were recorded on fifteen variables (standardized test scores) for 331 subjects between the ages of five years, 0 months and eleven years, eleven months. From an obtained correlational matrix, a list was constructed of all possible three-variable combinations.

In order to identify those three-variable combinations which would theoretically result in an increase or in no change in the correlational values existing between variables y and z in the restricted samples and the unrestricted population, formula (1) was manipulated so that the functional values are the correlational values which exist in the unrestricted population. Given such a manipulation, the prediction of the correlational value of y and z in the restricted sample (r'_{yz}) is found in (4).

$$r'_{yz} = \frac{R_{yz} + K'R_{xy}R_{xz}}{\sqrt{1 + K'^2R_{xy}^2} \sqrt{1 + K'^2R_{xz}^2}} \quad (4)$$

where

- r'_{yz} = the predicted correlational value of y and z in the restricted sample,
- R_{yz} = the correlational value of y and z in the unrestricted population,
- R_{xy} = the correlational value of x and y in the unrestricted population,
- R_{xz} = the correlational value of x and z in the unrestricted population, and
- K' = $\frac{\sigma_x^2}{\sigma_{x'}^2} - 1$

Employing formula (4) with $K' = -.5$ (a selection ratio of 2), a computer analysis revealed several three-variable combinations whose

values, when substituted into formula (4), produced the inequality $r'_{yz} > R_{yz}$. From this group of possible combinations, four three-variable combinations were selected on the basis of the largest number of subjects available. These four groups were treated as unrestricted populations, each consisting of 206 subjects.

Group I data consisted of scores on the following variables: standard scores on the Word Recognition Subtest of the Wide Range Achievement test (WRAT-R), the Social Age Score of the Vineland Social Maturity Scale (SA), and the Wechsler Intelligence Scale for Children full scale I.Q. score (WISC-F). The correlational values for these variables are found in Table 1.

Table 1
Group I Correlation Matrix for SA, WISC-F and WRAT-R
Scores in the Unrestricted Population

Variable	x	y	z
x SA	1.0000	.3417	-.0385
y WISC-F		1.0000	.5160
z WRAT-R			1.0000

Group II data consisted of scores on the following variables: standard scores on the WRAT-R, SA, and the Wechsler Intelligence Scale for Children Verbal Scale I.Q. Score (WISC-V). The correlational values for these variables are found in Table 2.

Table 2

Group II Correlation Matrix for SA, WISC-V and WRAT-R
Scores in the Unrestricted Population

Variable	x	y	z
x SA	1.0000	.2858	-.0385
y WISC-V		1.0000	.5333
z WRAT-R			1.0000

Group III data consisted of scores on the following variables: standard scores on the Spelling Subtest of the Wide Range Achievement Test (WRAT-S), SA and the WISC-F. The correlational values for these variables are found in Table 3.

Table 3

Group III Correlation Matrix for SA, WISC-F and WRAT-S
Scores in the Unrestricted Population

Variable	x	y	z
x SA	1.0000	.3417	-.0083
y WISC-F		1.0000	.5227
z WRAT-S			1.0000

Group IV data consisted of scores on the following variables: standard scores on the WRAT-S, WISC-V, and SA. The correlational values for these variables are found in Table 4.

Table 4

Group IV Correlation Matrix for SA, WISC-V and WRAT-S
Scores in the Unrestricted Population

Variable	x	y	z
x SA	1.0000	.2858	- .0083
y WISC-V		1.0000	.4957
z WRAT-S			1.0000

Restriction of range was produced on variable x (SA) in all four groups so that a selection ratio of 1.995 was obtained. In each group restriction was performed by consecutively deleting the lowest scores until the desired ratio was reached. This method of selection was employed because it is typical of the kind which occurs in practical situations. Each of the four resulting restricted samples contained 130 subjects. The correlational values in the restricted samples are found in Tables 5, 6, 7, and 8.

Table 5

Group I Correlation Matrix for SA, WISC-F, and WRAT-R
Scores in the Restricted Sample

Variable	x	y	z
x SA	1.0000	.0504	- .1868
y WISC-F		1.0000	.5697
z WRAT-R			1.0000

Table 6

Group II Correlation Matrix for SA, WISC-V, and WRAT-R
Scores in the Restricted Sample

Variable	x	y	z
z SA	1.0000	-.0100	-.1868
y WISC-V		1.0000	.6363
z WRAT-R			1.0000

Table 7

Group III Correlation Matrix for SA, WISC-F, and WRAT-S
Scores in the Restricted Sample

Variable	x	y	z
x SA	1.0000	.0504	-.1967
y WISC-F		1.0000	.4772
z WRAT-S			1.0000

Table 8

Group IV Correlation Matrix for SA, WISC-V, and WRAT-S
Scores in the Restricted Sample

Variable	x	y	z
x SA	1.0000	-.0100	-.1967
y WISC-V		1.0000	.5014
z WRAT-S			1.0000

With the values contained in the eight correlation matrices R'_{yz} and r'_{yz} values were computed by employing formulas (1) and (4). The values for R'_{yz} , r'_{yz} , R_z and r_{yz} for all four groups are presented in Table 9.

Table 9

Observed and Predicted Correlational Values for Variable y and z in the Unrestricted Population and in the Restricted Sample for Groups I - IV

Group	R_{yz}	r_{yz}	R'_{yz}	r'_{yz}
I	.5160	.5697	.5502	.5387
II	.5333	.6363	.6273	.5504
III	.5227	.4772	.4579	.5401
IV	.4957	.5014	.4799	.4996

A Z-test for correlational values (Edwards, 1967) was used to test for the significance of differences between R_{yz} and r_{yz} , R_{yz} and R'_{yz} , and r_{yz} and r'_{yz} in each group. Employing a non-directional test the levels of probability associated with the twelve obtained values of Z were all greater than .05, i.e., there were no significant differences between R_{yz} and r_{yz} , R_{yz} and R'_{yz} and r_{yz} and r'_{yz} within each of the four groups.

Discussion and Conclusions

The problem investigated in this study was twofold. First, an attempt was made to determine whether or not correlational values obtained from real data increase or remain unchanged as a result of restriction of range under those conditions outlined by Levin (1972)

and Elshout and Roe (1973). Secondly, an attempt was made to assess the accuracy of the case III correction formula under conditions of extreme univariate selection.

Regarding the first question, results indicate that in each of the four groups the two relevant correlational values (R_{yz} , r_{yz}) were not significantly different. This finding lends empirical support for the assertions of Levin (1972) and Elshout and Roe (1973), and is contradictory to traditional theory. The author therefore concludes that restriction of range in the three-variable case does not invariably decrease the correlational value between incidental variables.

Regarding the second question, the results indicate that in each of the four groups the two relevant correlational values (R_{yz} , R'_{yz}) were not significantly different. This finding lends empirical support for the accuracy and appropriateness of the case III correction formula under conditions of extreme univariate selection.

The present research seems to clarify some of the issues involved in accurately inferring parameters within the correlation model. This study also supports the contention of Levin (1972) that correlations in selected samples are not always underestimated values. In the future, researchers should be more cautious in citing restriction as the "cause" of observed values smaller than those expected. Additionally, researchers should take note that the conditions necessary for the presence of this phenomenon apparently occur in real data with sufficient regularity to necessitate a check on one's values, with the correction formula, in order to determine the accuracy of sample statistics.

REFERENCES

- Edwards, A.L. Statistical methods (2nd ed.) New York: Holt, Rinehart & Winston, 1967.
- Elshout, J.J., & Roe, R.A. Restriction of range in the population. Educational and Psychological Measurement, 1973, 33, 53-62.
- Gullford, J.P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1956.
- Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.
- Levin, J. The occurrence of an increase in correlation by restriction of range. Psychometrika, 1972, 37, 93-97.
- Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.
- Pearson, K. Mathematical contributions to the theory of education--XI. On the influence of natural selection on the variability and correlation of organs. Transatlantic Royal Society, 1903, 200, 1-66. Cited by S. Rydberg, Bias in prediction. Stockholm: Almqvist & W. Ksell, 1963, p. 25.
- Peters, C.C., & Van Voorhis, W.R. Statistical procedures and their mathematical basis. New York: McGraw-Hill, 1940.
- Rydberg, S. Bias in prediction. Stockholm: Almqvist & W. Ksell, 1963.
- Thorndike, R.L. Personnel selection. New York: Wiley, 1949.